A METHOD AND MECHANISM FOR THE CREATION, MAINTENANCE, AND COMPARISON OF SEMANTIC ABSTRACTS

RELATED APPLICATION DATA

This application is related to co-pending U.S. Patent application Serial No. 09/109,804, titled "METHOD AND APPARATUS FOR SEMANTIC CHARACTERIZATION," filed July 2, 1998, and to co-pending U.S. Patent application Serial No. 09/512,963, titled "CONSTRUCTION, MANIPULATION, AND COMPARISON OF A MULTI-DIMENSIONAL SEMANTIC SPACE," filed February 25, 2000.

FIELD OF THE INVENTION

This invention pertains to determining the semantic content of documents, and more particularly to summarizing and comparing the semantic content of documents to determine similarity.

BACKGROUND OF THE INVENTION

U.S. Patent application Serial No. 09/512,963, titled "CONSTRUCTION, MANIPULATION, AND COMPARISON OF A MULTI-DIMENSIONAL SEMANTIC SPACE," filed February 25, 2000, describes a method and apparatus for mapping terms in a document into a topological vector space. Determining what documents are about requires interpreting terms in the document through their context. For example, whether a document that includes the word "hero" refers to sandwiches or to a person of exceptional courage or strength is determined by context. Although taking a term in the abstract will generally not give the reader much information about the content of a document, taking several important terms will usually be helpful in determining content.

The content of documents is commonly characterized by an abstract that provides a high-level description of the contents of the document and provides the reader with some expectation of what may be found within the contents of the document. (In fact, a single

document can be summarized by multiple different abstracts, depending on the context in which the document is read.) Patents are a good example of this commonly used mechanism. Each patent is accompanied by an abstract that provides the reader with a description of what is contained within the patent document. However, each abstract must be read and compared by a cognitive process (usually a person) to determine if various abstracts might be describing content that is semantically close to the research intended by the one searching the abstracts.

Accordingly, a need remains for a way to associate semantic meaning to documents using dictionaries and bases, and for a way to search for documents with content similar to a given document, both generally without requiring user involvement.

SUMMARY OF THE INVENTION

To determine a semantic abstract for a document, the document is parsed into phrases. The phrases can be drawn from the entire document, or from only a portion of the document (e.g., an abstract). State vectors in a topological vector space are constructed for each phrase in the document. The state vectors are collected to form the semantic abstract. The state vectors can also be filtered to reduce the number of vectors comprising the semantic abstract. Once the semantic abstract for the document is determined, the semantic abstract can be compared with a semantic abstract for a second document to determine how similar their contents are. The semantic abstract can also be compared with other semantic abstracts in the topological vector space to locate semantic abstracts associated with other documents with similar contents.

The foregoing and other features, objects, and advantages of the invention will become more readily apparent from the following detailed description, which proceeds with reference to the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 shows a two-dimensional topological vector space in which state vectors are used to determine a semantic abstract for a document.

FIG. 2 shows a two-dimensional topological vector space in which semantic abstracts for two documents are compared by measuring the Hausdorff distance between the semantic abstracts.

FIG. 3 shows a two-dimensional topological vector space in which the semantic abstracts for the documents of FIG. 2 are compared by measuring the angle and/or distance between centroid vectors for the semantic abstracts.

FIG. 4 shows a computer system on which the invention can operate to construct semantic abstracts.

FIG. 5 shows a computer system on which the invention can operate to compare the semantic abstracts of two documents.

FIG. 6 shows a flowchart of a method to determine a semantic abstract for a document in the system of FIG. 4 by extracting the dominant phrases from the document.

FIG. 7 shows a flowchart of a method to determine a semantic abstract for a document in the system of FIG. 4 by determining the dominant context of the document.

FIG. 8 shows a dataflow diagram for the creation of a semantic abstract as described in FIG. 7.

FIG. 9 shows a flowchart showing detail of how the filtering step of FIG. 7 can be performed.

FIG. 10 shows a flowchart of a method to compare two semantic abstracts in the system of FIG. 5.

FIG. 11 shows a flowchart of a method in the system of FIG. 4 to locate a document with content similar to a given document by comparing the semantic abstracts of the two documents in a topological vector space.

FIG. 12 shows a saved semantic abstract for a document according to the preferred embodiment.

FIG. 13 shows a document search request according to the preferred embodiment.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

Determining Semantic Abstracts

A semantic abstract representing the content of the document can be constructed as a set of vectors within the topological vector space. (The construction of state vectors in a topological vector space is described in U.S. Patent application Serial No. 09/512,963, titled "CONSTRUCTION, MANIPULATION, AND COMPARISON OF A MULTI-DIMENSIONAL SEMANTIC SPACE," filed February 25, 2000, incorporated by reference

herein and referred to as "the Construction application.") FIG. 1 shows a two-dimensional topological vector space in which state vectors are used to construct a semantic abstract for a document. (FIG. 1 and FIGs. 2 and 3 to follow, although accurate representations of a topological vector space, are greatly simplified for example purposes, since most topological vector spaces will have significantly higher dimensions.) In FIG. 1, the "x" symbols locate the heads of state vectors for terms in the document. (For clarity, the line segments from the origin of the topological vector space to the heads of the state vectors are not shown in FIG. 1.) Semantic abstract 105 includes a set of vectors for the document. As can be seen, most of the state vectors for this document fall within a fairly narrow area of semantic abstract 105. Only a few outliers fall outside the main part of semantic abstract 105.

Now that semantic abstracts have been defined, two questions remain: what words are selected to be mapped into state vectors in the semantic abstract, and how is distance measured between semantic abstracts. The first question will be put aside for the moment and returned to later.

Revisiting Semantic Distance

Recall that in the Construction application it was shown that $\mathcal{H}(\mathbf{S})$ is the set of all compact (non-empty) subsets of a metrizable space $\mathbf{S}$. The Hausdorff distance h is defined as follows: Define the pseudo-distance $\xi(x, u)$ between the point $x \in \mathbf{S}$ and the set $u \in \mathcal{H}(\mathbf{S})$ as

$$\xi(x, u) = \min\{d(x, y) : y \in u\}.$$

Using ξ define another pseudo-distance $\lambda(u, v)$ from the set $u \in \mathcal{H}(\mathbf{S})$ to the set $v \in \mathcal{H}(\mathbf{S})$:

$$\lambda(u, v) = \max\{\xi(x, v) : x \in u\}.$$

Note that in general it is *not* true that $\lambda(u, v) = \lambda(v, u)$. Finally, define the distance $h(u, v)$ between the two sets $u, v \in \mathcal{H}(\mathbf{S})$ as

$$h(u, v) = \max\{\lambda(u, v), \lambda(v, u)\}.$$

The distance function h is called the *Hausdorff* distance. Note that

$h(u, v) = h(v, u)$,

$0 < h(u, v) < \infty$ for all $u, v \in \mathcal{H}(\mathbf{S})$, $u \neq v$,

$h(u, u) = 0$ for all $u \in \mathcal{H}(\mathbf{S})$, and

$h(u, v) \leq h(u, w) + h(w, v)$ for all $u, v, w \in \mathcal{H}(\mathbf{S})$.

Measuring Distance between Semantic Abstracts

If $\mathcal{H}(\mathbf{S})$ is the topological vector space and u and v are semantic abstracts in the topological vector space, then Hausdorff distance function h provides a measure of the distance between semantic abstracts. FIG. 2 shows a two-dimensional topological vector space in which semantic abstracts for two documents are compared. (To avoid clutter in the drawing, FIG. 2 shows the two semantic abstracts in different graphs of the topological vector space. The reader can imagine the two semantic abstracts as being in the same graph.) In FIG. 2, semantic abstracts 105 and 205 are shown. Semantic abstract 105 can be the semantic abstract for the known document; semantic abstract 205 can be a semantic abstract for a document that may be similar to the document associated with semantic abstract 105. Using the Hausdorff distance function h, the distance 210 between semantic abstracts 105 and 205 can be quantified. Distance 210 can then be compared with a classification scale to determine how similar the two documents are.

Although the preferred embodiment uses the Hausdorff distance function h to measure the distance between semantic abstracts, a person skilled in the art will recognize that other distance functions can be used. For example, FIG. 3 shows two alternative distance measures for semantic abstracts. In FIG. 3, the semantic abstracts 105 and 205 have been reduced to a single vector. Centroid 305 is the center of semantic abstract 105, and centroid 310 is the center of semantic abstract 205. (Centroids 305 and 310 can be defined using any measure of central tendency.) The distance between centroids 305 and 310 can be measured directly as distance 315, or as angle 320 between the centroid vectors.

As discussed in the Construction application, different dictionaries and bases can be used to construct the state vectors. It may happen that the state vectors comprising each semantic abstract are generated in different dictionaries or bases and therefore are not directly comparable. But by using a topological vector space transformation, the state vectors for one of the semantic abstracts can be mapped to state vectors in the basis for the other semantic

abstract, allowing the distance between the semantic abstracts to be calculated. Alternatively, each semantic abstract can be mapped to a normative, preferred dictionary/basis combination.

Which Words?

Now that the question of measuring distances between semantic abstracts has been addressed, the question of selecting the words to map into state vectors for the semantic abstract can be considered.

In one embodiment, the state vectors in semantic abstract 105 are generated from all the words in the document. Generally, this embodiment will produce a large and unwieldy set of state vectors. The state vectors included in semantic abstract 105 can be filtered from the dominant context. A person skilled in the art will recognize several ways in which this filtering can be done. For example, the state vectors that occur with the highest frequency, or with a frequency greater than some threshold frequency, can be selected for semantic abstract 105. Or those state vectors closest to the center of the set can be selected for semantic abstract 105. Other filtering methods are also possible. The set of state vectors, after filtering, is called the *dominant vectors*.

In another embodiment, a phrase extractor is used to examine the document and select words representative of the context of the document. These selected words are called *dominant phrases*. Typically, each phrase will generate more than one state vector, as there are usually multiple lexemes in each phrase. But if a phrase includes only one lexeme, it will map to a single state vector. The state vectors in semantic abstract 105 are those corresponding to the selected dominant phrases. The phrase extractor can be a commercially available tool or it can be designed specifically to support the invention. Only its function (and not its implementation) is relevant to this invention. The state vectors corresponding to the dominant phrases are called *dominant phrase vectors*.

The semantic abstract is related to the level of abstraction used to generate the semantic abstract. A semantic abstract that includes more detail will generally be larger than a semantic abstract that is more general in nature. For example, an abstract that measures to the concept of "person" will be smaller and more abstract than one that measures to "man," "woman," "boy," "girl," etc. By changing the selection of basis vectors and/or dictionary

when generating the semantic abstract, the user can control the level of abstraction of the semantic abstract.

Despite the fact that different semantic abstracts can have different levels of codified abstraction, the semantic abstracts can still be compared directly by properly manipulating the dictionary (topology) and basis vectors of each semantic space being used. All that is required is a topological vector space transformation to a common topological vector space. Thus, semantic abstracts that are produced by different authors, mechanisms, dictionaries, etc. yield to comparison via the invention.

Systems for Building and Using Semantic Abstracts

FIG. 4 shows a computer system 405 on which a method and apparatus for using a multi-dimensional semantic space can operate. Computer system 405 conventionally includes a computer 410, a monitor 415, a keyboard 420, and a mouse 425. But computer system 405 can also be an Internet appliance, lacking monitor 415, keyboard 420, or mouse 425. Optional equipment not shown in FIG. 4 can include a printer and other input/output devices. Also not shown in FIG. 4 are the conventional internal components of computer system 405: e.g., a central processing unit, memory, file system, etc.

Computer system 405 further includes software 430. In FIG. 4, software 430 includes phrase extractor 435, state vector constructor 440, and collection means 445. Phrase extractor 435 is used to extract phrases from the document. Phrases can be extracted from the entire document, or from only portions (such as one of the document's abstracts or topic sentences of the document). Phrase extractor 435 can also be a separate, commercially available piece of software designed to scan a document and determine the dominant phrases within the document. Commercially available phrase extractors can extract phrases describing the document that do not actually appear within the document. The specifics of how phrase extractor 435 operates are not significant to the invention: only its function is significant. Alternatively, phrase extractor can extract all of the words directly from the document, without attempting to determine the "important" words.

State vector constructor 440 takes the phrases determined by phrase extractor 435 and constructs state vectors for the phrases in a topological vector space. Collection means 445 collects the state vectors and assembles them into a semantic abstract.

Computer system 405 can also include filtering means 450. Filtering means 450 reduces the number of state vectors in the semantic abstract to a more manageable size. In the preferred embodiment, filtering means 450 produces a model that is distributed similarly to the original state vectors in the topological vector space: that is, the probability distribution function of the filtered semantic abstract should be similar to that of the original set of state vectors.

It is possible to create semantic abstracts using both commercially available phrase extractors and the words of the document. When both sources of phrases are used, filtering means 450 takes on a slightly different role. First, since there are three sets of state vectors involved (those generated from phrase extractor 435, those generated from the words of the document, and the final semantic abstract), terminology can be used to distinguish between the two results. As discussed above, the phrases extracted by the commercially available phrase extractor are called *dominant phrases*, and the state vectors that result from the dominant phrases are called *dominant phrase vectors*. The state vectors that result from the words of the document are called *dominant vectors*. Filtering means 450 takes both the dominant phrase vectors and the dominant vectors, and produces a set of vectors that constitute the semantic abstract for the document. This filtering can be done in several ways. For example, the dominant phrase vectors can be reduced to those vectors with the highest frequency counts within the dominant phrase vectors. The filtering can also reduce the dominant vectors based on the dominant phrase vectors. The dominant vectors and the dominant phrase vectors can also be merged into a single set, and that set reduced to those vectors with the greatest frequency of occurrence. A person skilled in the art will also recognize other ways the filtering can be done.

Although the document operated on by phrase extractor 435 can be found stored on computer system 405, this is not required. FIG. 4 shows computer system 405 accessing document 460 over network connection 465. Network connection 465 can include any kind of network connection. For example, network connection 465 can enable computer system 405 to access document 460 over a local area network (LAN), a wide area network (WAN), a global internetwork, or any other type of network. Similarly, once collected, the semantic abstract can be stored somewhere on computer system 405, or can be stored elsewhere using network connection 465.

FIG. 5 shows computer system 405 equipped with software 505 to compare semantic abstracts for two documents. Software 505 includes semantic abstracts 510 and 515 for the two documents being compared, measurement means 520 to measure the distance between the two semantic abstracts, and classification scale 525 to determine how "similar" the two semantic abstracts are.

Procedural Implementation

FIG. 6 is a flowchart of a method to construct a semantic abstract for a document in the system of FIG. 4 based on the dominant phrase vectors. At step 605, phrases (the dominant phrases) are extracted from the document. As discussed above, the phrases can be extracted from the document using a phrase extractor. At step 610, state vectors (the dominant phrase vectors) are constructed for each phrase extracted from the document. As discussed above, there can be more than one state vector for each dominant phrase. At step 615, the state vectors are collected into a semantic abstract for the document.

Note that phrase extraction (step 605) can be done at any time before the dominant phrase vectors are generated. For example, phrase extraction can be done when the author generates the document. In fact, once the dominant phrases have been extracted from the document, creating the dominant phrase vectors does not require access to the document at all. If the dominant phrases are provided, the dominant phrase vectors can be constructed without any access to the original document.

FIG. 7 is a flowchart of a method to construct a semantic abstract for a document in the system of FIG. 4 based on the dominant vectors. At step 705, words are extracted from the document. As discussed above, the words can be extracted from the entire document or only portions of the document (such as one of the abstracts of the document or the topic sentences of the document). At step 710, a state vector is constructed for each word extracted from the document. At step 715, the state vectors are filtered to reduce the size of the resulting set, producing the dominant vectors. Finally, at step 720, the filtered state vectors are collected into a semantic abstract for the document.

As also shown in FIG. 7, two additional steps are possible, and are included in the preferred embodiment. At step 725, the semantic abstract is generated from both the dominant vectors and the dominant phrase vectors. As discussed above, the semantic abstract

can be generated by filtering the dominant vectors based on the dominant phrase vectors, by filtering the dominant phrase vectors based on the dominant vectors, or by combining the dominant vectors and the dominant phrase vectors in some way. Finally, at step 730, the lexeme and lexeme phrases corresponding to the state vectors in the semantic abstract are determined. Since each state vector corresponds to a single lexeme or lexeme phrase in the dictionary used, this association is easily accomplished.

As discussed above regarding phrase extraction in FIG. 6, the dominant vectors and the dominant phrase vectors can be generated at any time before the semantic abstract is created. Once the dominant vectors and dominant phrase vectors are created, the original document is not required to construct the semantic abstract.

FIG. 8 shows a dataflow diagram showing how the flowcharts of FIGs. 6 and 7 operate on document 460. Operation 805 corresponds to FIG. 6. Phrases are extracted from document 460, which are then processed into dominant phrase vectors. Operation 810 corresponds to steps 705, 710, and 715 from FIG. 7. Words in document 460 are converted and filtered into dominant vectors. Finally, operation 815 corresponds to steps 720, 725, and 730 of FIG. 7. The dominant phrase vectors and dominant vectors are used to produce the semantic abstract and the corresponding lexemes and lexeme phrases.

FIG. 9 shows more detail as to how the dominant vectors are filtered in step 715 of FIG. 7. As shown by step 905, the state vectors with the highest frequencies can be selected. Alternatively, as shown by steps 910 and 915, the centroid of the set of state vectors can be located, and the vectors closest to the centroid can be selected. (As discussed above, any measure of central tendency can be used to locate the centroid.) A person skilled in the art will also recognize other ways the filtering can be performed.

FIG. 10 is a flowchart of a method to compare two semantic abstracts in the system of FIG. 5. At step 1005 the semantic abstracts for the documents are determined. At step 1010, the distance between the semantic abstracts is measured. As discussed above, distance can be measured using the Hausdorff distance function h. Alternatively, the centroids of the semantic abstracts can be determined and the distance or angle measured between the centroid vectors. Finally, at step 1015, the distance between the state vectors is used with a classification scale to determine how closely related the contents of the documents are.

As discussed above, the state vectors may have been generated using different dictionaries or bases. In that case, the state vectors cannot be compared without a topological vector space transformation. This is shown in step 1020. After the semantic abstracts have been determined and before the distance between them is calculated, a topological vector space transformation can be performed to enable comparison of the semantic abstracts. One of the semantic abstracts can be transformed to the topological vector space of the other semantic abstract, or both semantic abstracts can be transformed to a normative, preferred basis.

FIG. 11 is a flowchart of a method to search for documents with semantic abstracts similar to a given document in the system of FIG. 5. At step 1105, the semantic abstract for the given document is determined. At step 1110, a second document is located. At step 1115, a semantic abstract is determined for the second document. At step 1120, the distance between the semantic abstracts is measured. As discussed above, the distance is preferably measured using the Hausdorff distance function h, but other distance functions can be used. At step 1130, the distance between the semantic abstracts is used to determine if the documents are similar. If the semantic abstracts are similar, then at step 1135 the second document is selected. Otherwise, at step 1140 the second document is rejected.

Whether the second document is selected or rejected, the process can end at this point. Alternatively, the search can continue by returning to step 1110, as shown by dashed line 1145. If the second document is selected, the distance between the given and second documents can be preserved. The preserved distance can be used to rank all the selected documents, or it can be used to filter the number of selected documents. A person skilled in the art will also recognize other uses for the preserved distance.

Note that, once the semantic abstract is generated, it can be separated from the document. Thus, in FIG. 11, step 1105 may simply include loading the saved semantic abstract. The document itself may not have been loaded or even may not be locatable. FIG. 12 shows saved semantic abstract 1202 for a document. In FIG. 12, semantic abstract 1202 is saved; the semantic abstract can be saved in other formats (including proprietary formats). Semantic abstract 1202 includes document reference 1205 from which the semantic abstract was generated, vectors 1210 comprising the semantic abstract, and dictionary reference 1215

and basis reference 1220 used to generate vectors 1210. Document reference 1205 can be omitted when the originating document is not known.

FIG. 13 shows document search request 1302. Document search request 1302 shows how a search for documents with content similar to a given document can be formed. Document search request 1302 is formed using HTTP, but other formats can be used. Document search request 1302 includes list 1305 of documents to search, vectors 1310 forming the semantic abstract, dictionary reference 1315 and basis reference 1320 used to generate vectors 1310, and acceptable distances 1325 for similar documents. Note that acceptable distances 1325 includes both minimum and maximum acceptable distances. But a person skilled in the art will recognize that only a minimum or maximum distance is necessary, not both.

Having illustrated and described the principles of our invention in a preferred embodiment thereof, it should be readily apparent to those skilled in the art that the invention can be modified in arrangement and detail without departing from such principles. We claim all modifications coming within the spirit and scope of the accompanying claims.